

Face Identification with Second-Order Pooling

Fumin Shen, Chunhua Shen and Heng Tao Shen

Abstract

Automatic face recognition has received significant performance improvement by developing specialised facial image representations. On the other hand, generic object recognition has rarely been applied to the face recognition. Spatial pyramid pooling of features encoded by an over-complete dictionary has been the key component of many state-of-the-art image classification systems. Inspired by its success, in this work we develop a new face image representation method inspired by the second-order pooling in [1], which was originally proposed for image segmentation.

The proposed method differs from the previous methods in that, we encode the densely extracted local patches by a small-size dictionary; and the facial image signatures are obtained by pooling the second-order statistics of the encoded features. We show the importance of pooling on encoded features, which is bypassed by the original second-order pooling method to avoid the high computational cost. Equipped with a simple linear classifier, the proposed method outperforms the state-of-the-art face identification performance by large margins. For example, on the LFW databases, the proposed method performs better than the previous best by around 13% accuracy.

I. INTRODUCTION

Face identification aims to find the subject in the gallery most similar to the probe face image. Despite decades of research effort, it is still an active topic in computer vision due to both its wide applications and technical challenges. The challenges are typically caused by various intra-class variations (e.g., face expressions, poses, ages, image contaminations, etc.), or lack of sufficient training data [2]. One of the key problems is to generate a robust and discriminant representation for facial images. Extensive research effort in the literature has been devoted to projecting the face vectors to a low-dimensional subspace, e.g., as in the method of eigenfaces [3], Fisher-faces [4], Laplacian faces [5], etc. However, these holistic feature based methods often are incapable to cope with the aforementioned problems well.

Recently, sparse representation based face classification has achieved promising results [6], [7]. Different from previous methods, these methods compute the representation of the probe image to achieve the minimum representation error in terms of a set of training samples or a dictionary learned from training images. Many algorithms have been developed in this category, which achieve state-of-the-art performance on face recognition with image corruptions [6], face disguises [7], [8] and small-size training data [9], [10].

To improve the face recognition performance, many local feature based methods has been proposed, which tend to show superior results over those based holistic features. Typical methods in this group include histograms of local binary patterns (LBP) [11], histograms of various Gabor features [12], [13], [14] and their fusions [15]. These local feature based methods have been proven to be more robust to mis-alignment and occlusions.

On the other hand, the local feature based image representation—bag-of-visual-words (BOV)—has been shown state-of-the-art recognition accuracy [16]. The typical pipeline of BOV is: low-level local feature extraction (raw pixels, SIFT etc.), feature quantization or encoding against a pre-trained dictionary, and descriptor generation by spatially pooling the encoded local features. This pipeline has been shown to achieve the state-of-the-art performance in generic image classification [17], [18], [19]. Despite the success of the BOV model in image classification, it has been rarely applied to face recognition.

The dimensionality of the learned image descriptor through the BOV pipeline is mainly determined by the size of trained dictionary (dimension of the encoded local features) and the pooling pyramid grids. It has been shown

F. Shen is with School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, P.R. China (e-mail: fumin.shen@gmail.com; shenhengtao@hotmail.com). Part of this work was done when the first author was visiting The University of Adelaide. Correspondence should be addressed to F. Shen.

C. Shen is with The Australian Center for Visual Technologies, and School of Computer Science at The University of Adelaide, SA 5005, Australia (e-mail: chunhua.shen@adelaide.edu.au).

H. T. Shen is with School of Information Technology and Electrical Engineering, The University of Queensland, Australia and School of Computer Science and Engineering, University of Electronic Science and Technology of China (E-mail: shenht@itee.uq.edu.au).

that a large dictionary size is critical to achieve a high accuracy for generic image classification [20]. In the meantime, pooling features over a spatial region leads to more compact representations, and also helps to make the representation invariant to image transformation and more robust [21]. The spatial pyramid pooling model [18] has made a remarkable success, for example, in conjunction with sparse coding techniques [19].

Average pooling and max-pooling are the two most popular pooling methods. The latter method usually leads to superior performance to the former one [19], [22]. Most previous methods compute first-order statistics in the pooling stage. In contrast, recently the average and max-pooling methods that incorporate the second-order information of local features have been proposed in [1] for image segmentation. Without an encoding stage, the second-order pooling strategy of [1] is directly applied to the raw SIFT descriptors.

Inspired by both the BOV model and the second-order pooling method of [1], here we propose a new method for facial image representation. First, local raw patches are densely extracted from the face images. The local patches are then encoded by an *small-size* dictionary, e.g., trained by K-means. The encoded features are finally pooled by employing the second-order statistics over a multi-level pyramid. The efficacy of the learned facial features are verified by the state-of-the-art performance on several public benchmark databases.

The BOV model has been less frequently studied on face recognition problems. The Fisher Vectors on densely sampled SIFT features were adopted in [23] for face verification problems. In contrast, we focus on raw intensity features. Also note that no pooling is applied in this method. Combination of sparse coding and spatial max-pooling has been used in [24] for face recognition. However, only the first-order statistics are computed in the pooling stage.

Our contributions mainly include:

1. We propose a new facial representation method based on a combination of the BOV model and second-order pooling. To our knowledge, this is the first face feature extraction method using the second-order pooling technique.
2. Different from the standard BOV methods which usually involve an over-complete dictionary, we show that, a very small number of dictionary basis are sufficient for face identification problems, in conjunction with the second-order pooling. In contrast to the method in [1], which does not apply encoding, we show that feature encoding is critically important for face identification and always improves the recognition accuracy.
3. Coupling with a simple linear classifier, the proposed method outperforms those state-of-the-art by large margins on several benchmark databases, including AR, FERET and LFW. In particular, the proposed method achieves perfect recognitions (100% accuracies) on the ‘Fb’ subset of the FERET dataset and the ‘sunglasses’ and ‘scarves’ subsets of the AR dataset. Our method obtains a higher accuracy than the best previous result by around 13% on LFW.

II. THE PROPOSED METHOD

In this section we present the details of the proposed method. We focus on extracting a discriminant representation of an image based on its raw intensity feature other than other specific designed ones like SIFT.

A. Dense local patch extraction

Without loss of generality, suppose that a facial image is of $d \times d$ pixels. As the first step, we extract overlapped local patches of size $r \times r$ pixels with a step of s pixels. Set $l = \lfloor \frac{d-r}{s} + 1 \rfloor$, then each image is divided into $l \times l$ patches. Let each local patch be a row vector \mathbf{x} .

It has been shown that dense feature extraction and the pre-processing step are critical for achieving better performance [20]. In practice, we extract local patches of 6×6 pixels with a stride of 1 pixel. We then perform normalization on \mathbf{x} as: $\hat{x}_i = (x_i - m)/v$, where x_i is the i^{th} element of x , and m and v are the mean and standard deviation of elements of \mathbf{x} . This operation contributes to local brightness and contrast normalization as in [20].

B. Unsupervised dictionary training

The goal of dictionary training is to generate a set of representative basis $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_i, \dots, \mathbf{d}_m\} \in \mathbf{R}^{d \times m}$. Here m is the number of atoms, and d is the input dimension. A great deal of unsupervised dictionary learning methods has been developed, for example the K-means clustering, sparse coding, K-SVD [25]. Dictionary can also be trained with the help of category information, such as the supervised sparse coding method [24]. We adopt the

K-means algorithm, since it is simple and effective. The dictionary size is a more important factor compared to the dictionary training algorithm.

With a first-order pooling method, it has been shown that the image classification accuracy is consistently improved as the dictionary size increases [20]. However, this is not necessarily true when a second-order pooling technique is applied. Perhaps surprisingly, with second-order pooling, a very small number of dictionary basis are sufficient to obtain high recognition accuracies. We will analysis this in the section III.

As a common pre-processing step in deep learning methods, whitening has been shown to yield sharply localized filters when dictionary are trained by clustering on raw data [20]. We apply the ZCA whitening on each patch [26] before the dictionary learning algorithm are applied.

C. Feature encoding

With the leaned dictionary, the pre-processed local patches are are then fed into the feature encoder to generate a set of mid-level features. Popular choices of encoding algorithms include the sparse coding [19], Locality-constrained linear coding (LLC [27]), etc. A good evaluation of different encoders can be found in [28]. The combinations of different dictionary learning and feature encoding methods are thoroughly studied in [29]. In our method, we adopt the soft threshold method, which encodes the patches by a simple feed-forward non-linearity with a fixed threshold. This simple encoder writes [29]:

$$\mathbf{f}_j(\mathbf{x}) = \max \{0, \mathbf{D}_j^\top \mathbf{x} - \alpha\}, \quad (1)$$

$$\mathbf{f}_{j+m}(\mathbf{x}) = \max \{0, -\mathbf{D}_j^\top \mathbf{x} - \alpha\}. \quad (2)$$

Here \mathbf{f}_j is the j^{th} entry of the encoded feature vector \mathbf{f} . Despite its simplicity, soft threshold achieves close performance with sparse coding on the image classification task.

D. Second-order pooling

As discussed before, feature pooling plays an important role in the BOV pipeline. The pooling procedure reduces the dimensionality of the learned mid-level features on each spatial region. Moreover, the pooled features are more robust to pose variations of face images.

Depart from most of the previous methods using the first-order pooling, following [1], we compute the second-order statistics of the encode features in the pooling stage. The second-order average-pooling over a spatial region R is defined as:

$$\mathbf{F}_{avg} = \sum_{i:\mathbf{f}_i \in R} \mathbf{f}_i \cdot \mathbf{f}_i^\top / |R|, \quad (3)$$

where \mathbf{f}_i is the column feature vector learned from region R and $|R|$ is the total number of feature vectors in region R . Through the outer product operation, information between all interacting pairs of descriptor dimensions is preserved. The computed \mathbf{F}_{avg} is a symmetric positive definite (SPD) matrix, which naturally forms a Riemannian manifold [1]. Mapping the second-order average pooling outputs by the Log-Euclidean metrics into the tangent space has been shown to significantly improve the classification. The power normalization [30] is also applied after the mapping in [1]. However, in our experiments, this operation do not shown any performance improvements for face identification. In practice, we only conduct the Log-Euclidean mapping:

$$\mathbf{F}_{avg}^{log} = \log(\mathbf{F}_{avg}). \quad (4)$$

The is computed by the algorithm in [31].

By concatenating together all the pooled second-order statistics over a multiple-level pyramid in one vector, we obtain the final representation of a face image. In this work, we feed the extracted features to a linear classifier to recognize the probe face image. A simple ridge regression based multi-class classifier was used in [32]. We use the same linear classifier for its computational efficiency. The ridge regression classifier has a closed-form solution, which makes the training even faster than the specialized linear support vector machine (SVM) solver LIBLINEAR [33]. Despite its simplicity, the classification performance of this ridge regression approach is on par with linear SVM [32]. Another benefit of this classifier is that, compared to the one-versus-all SVM, it only needs to compute the classification matrix \mathbf{W} once.

TABLE I

CROSS EVALUATION ACCURACIES (%) WITH VARYING POOLING PYRAMIDS VERSUS DICTIONARY SIZES. ALL THE REPORTED RESULTS IN THIS TABLE ARE BASED ON 5 INDEPENDENT DATA SPLITS. HERE WE VARY THE DEPTH OF PYRAMIDS FROM 3-LEVEL $\{1, 2, 4\}$ TO MAXIMUM 8-LEVEL $\{1, 2, 4, 6, 8, 10, 12, 15\}$.

pyramid levels	dictionary size						
	5	10	20	40	60	80	100
3	31.8 ± 4.1	76.5 ± 1.4	84.9 ± 1.3	86.2 ± 1.8	86.0 ± 1.6	85.2 ± 1.2	84.8 ± 1.5
4	66.1 ± 3.3	84.7 ± 1.7	88.0 ± 1.4	87.7 ± 2.6	86.7 ± 2.0	86.5 ± 1.9	86.1 ± 1.9
5	75.0 ± 2.0	86.5 ± 1.6	88.3 ± 1.3	88.2 ± 1.6	87.7 ± 1.3	88.0 ± 1.3	87.7 ± 1.1
6	77.0 ± 2.8	86.3 ± 1.7	88.3 ± 1.1	87.7 ± 1.4	-	-	-
8	76.5 ± 2.7	84.3 ± 1.8	86.1 ± 1.8	85.9 ± 1.7	-	-	-

III. DO WE NEED A LARGE DICTIONARY?

The key factors that affect the classification performance include the dictionary size and pooling pyramid levels. In addition, these two factors also determine the computational cost and the dimensionality of the learned image descriptor. In particular, the time complexity of the second-order pooling step is $O(m^2)$ with respect to the dictionary size m (the dimensionality of the encoded local feature).

In this section, we thoroughly evaluate the impact of the components of the proposed algorithms. We vary the dictionary size from 5 to 100, as shown in Table I. Different pooling pyramids are tested: from the 3-level pyramid $\{1, 2, 4\}$ to a maximum 8-level pyramid $\{1, 2, 4, 6, 8, 10, 12, 15\}$. With this 8-level pyramid, pooling is performed on regular grids of 1×1 , 2×2 , ..., 15×15 and the obtained pooled features are concatenated altogether. The evaluation is conducted on the LFW-a dataset [34], and all the images are down-sampled to 64×64 pixels. The dataset's description can be found in Section V. We set the number of training and testing samples per subject to 5 and 2, respectively. All the results reported in this section are based on 5 independent data splits.

As we can see from each row of Table I, the best recognition result is achieved with only a small number of dictionary basis for each pyramid. For example, the proposed algorithm reach its highest average recognition rate with a dictionary of only 20 atoms and a pyramid with 5 or 6 levels. This phenomenon is very different from the literature of generic classification, where the accuracy tends to be consistently improved as the dictionary size increases and a large over-complete dictionary is critical to achieve high accuracies [20]. A possible reason is that with second-order pooling, corresponding to all possible pairs of dictionary basis, the information between interacting pairs of local feature dimensions is preserved. This reduces the necessity of the use of redundant dictionary items.

From each column of Table I, we can observe that more pyramid levels (up to 6) tend to result in better performance. The information of the pooled representation is enriched from multi-scale patches. However redundant features do not always improve the final recognition rates. Accuracy drops are also observed with more than 8 pyramid levels, similar as the dictionary size.

Taken into account both the performance and computation cost, we set in the following paper the dictionary size and number of pyramid levels to 20 and 5, respectively.

IV. POOLING ON ENCODED FEATURES OR RAW PATCHES?

In the previous BOV methods using the first-order pooling, the encoding stage is usually in conjunction with an over-complete dictionary, which consequently produces high-dimensional local features. Different from the first-order pooling methods, second-order pooling results in much larger computational complexities due to the outer product operations. To avoid this, the method in [1] choose to directly apply pooling on the raw local descriptors (e.g., SIFT) without any encoding stage. The authors claim that good performance can be obtained without any feature coding due to that the preservation of the second-order statistics.

In this section, we will explore whether feature encoding is critical for face identification, based on raw local patches. Table II shows the identification rates of the second-order method with and without feature encoding on LFW and FERET. The number of training and testing samples are set to 5 and 2 respectively on both of the two databases. The datasets' description can be found in Section V.

It is clear that the classification accuracy is largely improved by second-order pooling encoded features than pooling raw features. In this case, the encoded features provide more discriminant information than the raw patches.

TABLE II

IDENTIFICATION ACCURACY (%) OF THE SECOND-ORDER METHOD WITH AND WITHOUT FEATURE ENCODING ON LFW AND FERET. A 5-LEVEL PYRAMID IS USED.

	with encoding	without encoding
LFW	88.3 ± 1.3	82.8 ± 2.4
FERET	98.5 ± 0.3	95.0 ± 1.0

Therefore, for face identification problems we suggest using the second-order pooling method on the encoded local features, generated from a dictionary which is not necessarily large.

V. EXPERIMENTAL RESULTS

In this section, we thoroughly evaluate the proposed method on several public facial image datasets including FERET [35], AR [36], LFW [37] and Pubfig83 [38]. Since there are a great deal of algorithms developed in the face identification community, we only compare our method with a few of them, which are representing or reported to achieve the state-of-the-art results. These methods include the superposed sparse representation classifier (SSRC [10]), local patch based Volterra [39], multi-scale patch based MSPCRC [40], and the popular BOV model Locality-constrained linear coding (LLC [27]). The soft thresholding (ST) method [29] with the first-order pooling is also taken into comparison. Both LLC and soft thresholding use a 3-level pyramid and 1024 dictionary size. We set the regularization parameter λ to 0.001 for SRC, RSC, MSPCRC and LLC, and 0.005 for SSRC, according to the authors' recommendation. The parameter α is set to 0.25 for ST. For fair comparisons, contrast normalization and ZCA whitening are performed for both LLC, ST and our method.

A. FERET

The FERET dataset [35] is a widely-used standard face recognition benchmark set provided by DARPA. Since the images in FERET are collected in multiple sessions during several years, they have complex intraclass variability. In our first experiment on FERET, we apply the standard protocols. With the gallery Fa (1196 images of 1196 subjects), we test on four probe sets: Fb (1195 images of 1195 subjects), Fc (194 images of 194 subjects), Duplicate I (722 images of 243 subjects, denoted as DupI), and Duplicated II (234 images of 75 subjects, denoted as DupII). Fb and Fc are captured with expression and illumination variations respectively. DupI and DupII are captured at different times. All the images are aligned based on the manually located eye centers and normalized to 150×130 pixels.

In this experiment, we mainly compare our facial image representation method to other feature extraction algorithms. The first five methods in Table III are specifically designed for face feature extraction, which are based on Gabor feature [13], [14] or feature fusion with different features feature [15], [41], [42]. LLC and ST are two BOV methods.

The compared results are shown in Table III. It is clear that, on all four sessions the proposed method achieves the best performance. Among the five specific facial representation methods, MBC obtains superior results on Fb and Fc, while Xie's method performs better on DupI and DupII. However, they are still inferior to our method, which is much simpler. We can also see that all the three BOV approaches LLC, ST and our method perform very well on the subset Fb and Fc with only expression and illumination variations. However, on the subset DupI and DupII with images taken at different years, LLC and ST obtain dramatic performance drop, while our second-order method still get very high accuracies. This demonstrates the discriminant ability of the proposed face image representation method.

In the second experiment on FERET, we used a subset of FERET which includes images with pose variations from 200 subjects. Each individual contains 7 samples with pose variations of up to 25 degrees. It is composed of images whose names are marked with 'ba', 'bj', 'bk', 'be', 'bf', 'bd' and 'bg'. These images are cropped and resized to 80×80 pixels [43]. We randomly select 5 samples for training and 2 samples for testing. The mean results of 5 independent runs with images down-sampled to 64×64 pixels are shown in Table IV.

We can clearly see that our method achieves the highest recognition rate. In particular, our method obtains an accuracy of 98.5% which is higher than the second best SSRC by 16.4%. RSC dose not show better results than SRC. The local patch based MSPCRC and Volterra do not perform well. This is probably because they are incapable of coping with pose variations.

TABLE III
RECOGNITION ACCURACIES (%) ON FERET WITH STANDARD PROTOCOLS. RESULTS OF THE FIRST 5 METHODS ARE CITED FROM [41]
WITH IDENTICAL SETTINGS.

Session	Fb(expression)	Fc (illumination)	DupI (aging)	DupII (aging)
HGPP [13]	97.5	99.5	79.5	77.8
Zou's method [14]	99.5	99.5	85.0	79.5
Tan's method [15]	98.0	98.0	90.0	85.0
Xie's method [42]	99.0	99.0	94.0	93.0
MBC [41]	99.7	99.5	93.6	91.5
LLC	99.6	100	78.4	69.2
ST	98.8	99.5	70.5	68.0
Ours	99.8	100	96.0	96.6

TABLE IV
RECOGNITION ACCURACIES (%) ON FERET WITH POSE VARIATIONS. RESULTS ARE BASED ON 5 INDEPENDENT RUNS.

Method	SRC	RSC	MSPCRC	Volterra	SSRC	Ours
Accuracy	73.8 ± 2.0	73.4 ± 2.6	42.3 ± 2.8	50.4 ± 2.3	82.1 ± 1.6	98.5 ± 0.3

B. AR

Since AR has been used very often to evaluate face recognition algorithms, we compare our methods to several published state-of-the-art results on this dataset. The AR dataset has 126 subjects (70 men and 56 women) and contains more than 4000 facial images. Each subject contains 26 images taken in two separate sessions. The images exhibit a number of variations including facial expression (neutral, smile, anger, scream), illumination (left light on, right light on, both sides light on) and occlusion (sunglasses, scarves). We select 100 subjects (50 men and 50 women) in our experiment. Four different situations are tested. For the ‘All’ situation, all the 13 images in the first session are used for training and the other 13 images in the second session for testing. For the ‘clean’ situation, 7 samples in each session with only illumination and expression changes are used for training and testing. For the ‘sunglasses’ and ‘scarf’ situation, 8 clean samples from two sessions are used for training and 2 images with sunglasses and scarf are used for testing. All the images are resized to 64×64 pixels. The test results are shown in Table V.

It is obvious that our method achieves the highest accuracies in all situations. In particular, our method achieves perfect recognitions (100%) when face images are with sunglasses and scarves occlusions on this dataset. To the best of our knowledge, our method is the only one achieving this results in both occlusion situations. Again, our method show its discriminant ability for face images.

C. LFW

Following the settings in [40], here we use LFW-a, an aligned version of LFW using commercial face alignment software [34]. A subset of LFW-a including 158 subjects with each subject more than 10 samples are used. The

TABLE V
COMPARISON WITH RECENT STATE-OF-THE-ART RESULTS (%) ON AR WITH FOUR DIFFERENT SETTINGS. FOR THE FIRST SIX METHODS, WE QUOTE THE BEST REPORTED RESULTS FROM THE CORRESPONDING PUBLISHED PAPERS, WITH THE SAME EXPERIMENT SETTING.

Situation	Clean	Sunglasses	Scarves	All
SRC [6]	92.9	87.0	57.5	-
RSC [8]	96.0	99.0	97.0	-
DRDL [44]	95.0	-	-	-
L_2 [45]	-	78.5	79.5	95.9
SSRC [10]	-	90.9	90.9	98.6
FDDL [7]	92.0	-	-	-
Volterra	90.9	96.1	92.1	87.5
ST	98.1	99.5	98.5	99.2
LLC	98.8	99.0	99.0	99.3
Ours	99.9	100	100	99.9

TABLE VI

RECOGNITION ACCURACY ON LFW WITH DOWNSAMPLED 32×32 IMAGES. RESULTS ARE BASED ON 20 INDEPENDENT RUNS. THE RESULTS FOR PNN, VOLTERRA AND MSPCRC ARE QUOTED FROM [40] WITH IDENTICAL SETTINGS.

Method	SRC	PNN[46]	Volterra	MSPCRC	ST	LLC	Ours
Accuracy	44.1 ± 2.6	47.4 ± 2.7	40.3 ± 2.7	49.0 ± 2.9	73.9 ± 1.6	64.1 ± 2.2	78.7 ± 3.1

TABLE VII

RECOGNITION ACCURACY (%) ON LFW-A WITH DIFFERENT IMAGE SIZES.

Method	32×32 pixels	64×64 pixels
MSPCRC	49.0 ± 2.9	47.7 ± 2.0
LLC	64.1 ± 2.2	75.4 ± 2.6
ST	73.9 ± 1.6	73.7 ± 2.0
Ours	78.7 ± 3.1	88.3 ± 1.3

images are cropped to 121×121 pixels. We randomly selected 5 samples for training and another 2 samples for testing. All the images are finally resized to 32×32 pixels as in [40].

Table VI lists the compared identification accuracies. Consistent with the previous results, our method outperforms all other algorithms by even larger gaps on this very challenging dataset. We obtain higher accuracies than MSPCRC and ST by 29.7% and 4.8%, respectively. Table VII shows the compared results with different image sizes. If we use a larger image size 64×64 , the accuracy of our method will increase to 88.3%, while ST and MSPCRC do not shown any improvements. LLC achieves better results with larger image dimension, however it is still inferior to our method.

D. Pubfig83

PubFig83 [38] is a subset of PubFig dataset [47], which include a large collection of real-world images of celebrities collected from the Internet. With 100 more samples in each of the 83 individual, PubFig83 is reconfigured for the problem of unconstrained face identification. We follow the evaluation protocol of [38] and use 90 samples for training, 10 samples for testing. The images are resized to 100×100 pixels and the rest results are based on 10 independent runs. The results are shown in Table VIII.

The first four results in Table VIII are quoted from [48] using different subspace method and combined representations of LBP, HOG and Gabor. RAW refers to the raw image representation, PLS refers to the partial least squares and PS-PLS refers to person specific PLS [48]. It is clear that our method outperforms all the other methods. In specific, the proposed method outperforms the best of other methods by 2.6%.

VI. CONCLUSION

In this paper, we propose a new representation method for facial images, inspired by the second-order pooling method in [1]. The key idea is to extract the second-order statistics of the encoded local features, which are generated by a small-size dictionary. We directly apply this method on densely extracted local patches other than specialized features like SIFT. We show that, with second-order pooling, the dictionary size (20 in all our experiments) is not necessarily as large as in the first-order method. We also show that the feature encoding procedure is critical for face identification problems, even if the dictionary is very small. The discriminant power of the proposed facial image representation methods has been verified by the state-of-the-art performance on several benchmark datasets. On the FERET database, for instance, our method achieves accuracies of 100%, 96.0% and 96.6% on the Fc, DupI and DupII subsets, respectively. The proposed method also outperforms the best report results on AR and LFW dataset on face identification problems.

TABLE VIII

RECOGNITION ACCURACY (%) ON PUBFIG83. THE FIRST FOUR METHODS USE COMBINED REPRESENTATIONS OF LBP, HOG AND GABOR [48] AND LINEAR SVMs.

Method	RAW	PCA	PLS	PS-PLS	Ours
Accuracy	82.6 ± 0.3	82.4 ± 0.3	83.0 ± 0.3	85.4 ± 0.3	88.0 ± 0.8

We plan to explore the efficacy of the proposed image representation method on face verification problems in the future work.

REFERENCES

- [1] Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Semantic segmentation with second-order pooling. In: Proc. Eur. Conf. Comp. Vis. Springer (2012) 430–443
- [2] Li, S.Z., Jain, A.K.: Handbook of Face Recognition. Springer London (2011)
- [3] Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neuroscience **3**(1) (January 1991) 71–86
- [4] Belhumeur, P.N., Hespanha, J.a.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Patt. Anal. Mach. Intell. **19** (July 1997) 711–720
- [5] He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face recognition using laplacianfaces. IEEE Trans. Patt. Anal. Mach. Intell. **27**(3) (march 2005) 328–340
- [6] Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31** (2009) 210–227
- [7] Yang, M., Zhang, L., Feng, X., Zhang, D.: Fisher discrimination dictionary learning for sparse representation. In: Proc. IEEE Int. Conf. Comp. Vis. (2011) 543–550
- [8] Yang, M., Zhang, L., Yang, J., Zhang, D.: Robust sparse coding for face recognition. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2011) 625–632
- [9] Deng, W., Hu, J., Guo, J.: Extended SRC: undersampled face recognition via intra-class variant dictionary. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9) (2012) 1864–1870
- [10] Deng, W., Hu, J., Guo, J.: In defense of sparsity based face recognition. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2013) 399–406
- [11] Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12) (2006) 2037–2041
- [12] Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R., Konen, W.: Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. Comput. **42**(3) (1993) 300–311
- [13] Zhang, B., Shan, S., Chen, X., Gao, W.: Histogram of gabor phase patterns (hgpp): a novel object representation approach for face recognition. IEEE Trans. Image Processing **16**(1) (2007) 57–68
- [14] Zou, J., Ji, Q., Nagy, G.: A comparative study of local matching approach for face recognition. IEEE Trans. Image Processing **16**(10) (2007) 2617–2628
- [15] Tan, X., Triggs, B.: Fusing gabor and LBP feature sets for kernel-based face recognition. In: International Workshop on Analysis and Modeling of Faces and Gestures. (2007) 235–249
- [16] Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. Volume 1. (2004) 1–2
- [17] Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: Proc. IEEE Conf. Comp. Vis. Volume 2. (2005) 1458–1465
- [18] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2006) 2169–2178
- [19] Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2009) 1794–1801
- [20] Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proc. Int. Conf. Artif. Intell. Stat. (2011) 215–223
- [21] Boureau, Y.L., Ponce, J., LeCun, Y.: A theoretical analysis of feature pooling in visual recognition. In: Proc. Int. Conf. Mach. Learn. (2010) 111–118
- [22] Boureau, Y., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2010) 2559–2566
- [23] Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: Proc. British Mach. Vis. Conf. Volume 1. (2013) 7
- [24] Yang, J., Yu, K., Huang, T.: Supervised translation-invariant sparse coding. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn., IEEE (2010) 3517–3524
- [25] Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Processing **54**(11) (2006) 4311–4322
- [26] Hyvarinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. **13**(4-5) (2000) 411–430
- [27] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2010) 3360–3367
- [28] Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: Proc. British Mach. Vis. Conf. (2011)
- [29] Coates, A., Ng, A.: The importance of encoding versus training with sparse coding and vector quantization. In: Proc. Int. Conf. Mach. Learn. (2011) 921–928
- [30] Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proc. Eur. Conf. Comp. Vis. Springer (2010) 143–156
- [31] Davies, P.I., Higham, N.J.: A schur-parlett algorithm for computing matrix functions. SIAM Journal on Matrix Analysis and Applications **25**(2) (2003) 464–485
- [32] Gong, Y., Lazebnik, S.: Comparing data-dependent and data-independent embeddings for classification and ranking of internet images. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2011) 2633–2640

- [33] Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: Liblinear: a library for large linear classification. *J. Mach. Learn. Res.* (2008) 1871–1874
- [34] Wolf, L., Hassner, T., Taigman, Y.: Similarity scores based on background samples. In: *Proc. Asian conf. Comp. Vis.* (2010) 88–97
- [35] Phillips, P., Wechsler, H., Huang, J., Rauss, P.: The FERET database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.* **16**(5) (1998) 295–306
- [36] Martinez, A., Benavente, R.: The AR Face Database. CVC, Tech. Rep. (1998)
- [37] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (October 2007)
- [38] Pinto, N., Stone, Z., Zickler, T., Cox, D.D.: Scaling-up Biologically-Inspired Computer Vision: A Case-Study on Facebook. In: *Workshop on Biologically Consistent Vision, IEEE Conf. Comp. Vis. Patt. Recogn.* (2011)
- [39] Kumar, R., Banerjee, A., Vemuri, B.C., Pfister, H.: Trainable convolution filters and their application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(7) (2012) 1423–1436
- [40] Zhu, P., Zhang, L., Hu, Q., Shiu, S.: Multi-scale patch based collaborative representation for face recognition with margin distribution optimization. In: *Proc. Eur. Conf. Comp. Vis.* (2012) 822–835
- [41] Yang, M., Zhang, L., Shiu, S.K., Zhang, D.: Monogenic binary coding: An efficient local feature extraction approach to face recognition. *IEEE Trans. Trans. Inf. Forensics Security* **7**(6) (2012) 1738–1751
- [42] Xie, S., Shan, S., Chen, X., Chen, J.: Fusing local patterns of gabor magnitude and phase for face recognition. *IEEE Trans. Image Processing* **19**(5) (2010) 1349–1361
- [43] Yang, J., Yang, J., Frangi, A.: Combined Fisherfaces framework. *Image Vis. Comput.* **21**(12) (2003) 1037–1044
- [44] Ma, L., Wang, C., Xiao, B., Zhou, W.: Sparse representation for face recognition based on discriminative low-rank dictionary learning. In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* (2012) 2586–2593
- [45] Shi, Q., Eriksson, A., van den Hengel, A., Shen, C.: Is face recognition really a compressive sensing problem? In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* (2011) 553–560
- [46] Kumar, R., Banerjee, A., Vemuri, B., Pfister, H.: Maximizing all margins: pushing face recognition with kernel plurality. In: *Proc. IEEE Int. Conf. Comp. Vis.* (2011) 2375–2382
- [47] Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: *Proc. IEEE Conf. Comp. Vis., IEEE* (2009) 365–372
- [48] Chiachia, G., Pinto, N., Schwartz, W.R., Rocha, A., Falcão, A.X., Cox, D.D.: Person-specific subspace analysis for unconstrained familiar face identification. In: *Proc. British Mach. Vis. Conf.* (2012) 1–12